Malcolm Gaynor and Parker Gibbons
STAT 306
Professor Hartlaub
Executive Summary

**MLB Roster Construction Based On Market Size**

We examined the financial metrics of all 30 MLB teams in order to determine whether market size had an impact on what different teams rely on to win and have playoff success. We also wanted to explore any possible similarities, including roster construction, between successful teams despite any differences in market size. To do this, we collected financial metrics for each MLB team over the last four seasons to classify different market sizes and collected data behind the dispersion of money for each team to see where each team allocated their payroll.

Our first step was using k-means clustering to classify each MLB team into a market size (small, medium, large). We wanted to know whether we could group up the teams by their financial metrics(total payroll, media market size, team valuation, and attendance and felt that K-means clustering was appropriate. The results of the clustering were pretty good as three different market sizes were clearly distinguished.

Our next step was creating and examining models to predict win percentage and playoff success created separately for teams of each market size. We first started with basic multiple linear regressions and found that each market size had different variables in their respective models used to predict winning and playoff success. In both cases of winning and playoff success, large market teams relied more on pitcher payroll and CBT space while medium market teams relied on pitcher payroll and outfield payroll, and small market teams relied on total payroll and the number of rookies on the roster. Notably, the models for the large markets explained more variation in both winning and playoff success than medium and small markets.

While these models explained a moderate to good amount of variability, we wanted to go more in-depth with modeling and explore some possible non-linear relationships and interaction among our predictor variables. This led us to use MARS (Multivariate Adaptive Regression Splines) which is a flexible nonparametric regression technique that fits piecewise linear functions to subsets of the data and offers interpretability and flexibility by capturing complex relationships between predictors and a response variable. This resulted in the same models as before now explaining more variation in both winning and playoff success while including less predictor variables, thus making the predictive power of our models stronger. These models revealed that small market teams rely on having less rookies and increased payroll for pitchers and outfielders. Medium market teams rely on having less rookies and increased payroll for infielders and catchers. Large market teams rely on signing more free agents and spending more money on pitchers.

Our work showed that MLB teams of different market sizes rely on different elements of roster construction to win and have success in the playoffs. Given more data on variables such as team revenue, contract lengths, merchandise sales and ticket prices, we could intensify our models and possibly explain more variation in both winning and playoff success. We could also use alternate clustering methods such as hierarchical clustering in order to classify market sizes. In addition, we could look into player performance metrics and possibly examine case studies of specific teams in a season. We could also add to our methodology by using degree 2 MARS for interaction terms. Moreover, our work showed the potential of examining how MLB teams of different financial circumstances approach winning and how they allocate the money they do have in order to have success.